ISOLATING THE EFFECTS OF INFLUENCIAL OBSERVATIONS

by

R. Dennis Cook

Technical Report # 283

Department of Applied Statistics
University of Minnesota
Saint Paul, MN 55108

March 18, 1977

SUMMARY

The distance measure proposed by Cook [2] for the detection of influential observations is extended to include subsets of the parameter vector in full rank linear regression models. An upper bound for the influence of an observation on all possible sets of linearly independent combinations of the elements of the parameter vector is given.

Key words: Influential observations, Confidence ellipsoids, Parameter subsets.

1. INTRODUCTION

Cook [2] proposed a measure based on confidence ellipsoids for judging the contribution of each data point to the determination of the least squares estimate, $\hat{\underset{\sim}{\beta}}$, of the parameter vector, $\underset{\sim}{\beta}$, in full rank linear regression models. The basic statistic, D_i, measures the change in $\hat{\underset{\sim}{\beta}}$ when the ith point is deleted. However, in many applications interest may center on some selected subset of $\underset{\sim}{\beta}$ rather than on the full vector. For example, in many analyses the constant term is of little or no interest and, thus, should be ignored when judging the influence of each data point. The generalized distance measure presented in [2] can be used to deal with such situations. However, the seemingly complicated form of this measure tends to overshadow the possibility of routine application. In this note we show that when subsets of $\underset{\sim}{\beta}$ are of interest the generalized measure reduces to an unexpectedly simple form. An example is presented. For completeness we first review the basic ideas.

Consider the full rank linear regression model

$$\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \qquad (1)$$

where $\underset{\sim}{Y}$ is an n×1 vector of observations, $\underset{\sim}{X}$ is an n×p full rank matrix of known constants, $\underset{\sim}{\beta}$ is a p×1 vector of unknown parameters and $\underset{\sim}{\varepsilon}$ is such that $E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$ and $V(\underset{\sim}{\varepsilon}) = \underset{\sim}{I}\,\sigma^2$.

Let $\underset{\sim}{A}$ denote a q×p rank q matrix,

$$\underset{\sim}{B} = \underset{\sim}{A}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{A}'$$

and

$$\underset{\sim}{\psi} = \underset{\sim}{A}\,\underset{\sim}{\beta}\,.$$

We assume that interest is in the q linearly independent combinations of the elements of $\underset{\sim}{\beta}$ specified by $\underset{\sim}{\psi}$.

The importance of the ith data point is judged by first computing the least squares estimate of $\underset{\sim}{\psi}$ with and without the point and, second, measuring the distance, $D_i(\underset{\sim}{\psi})$, between the two estimates as a monotonic function of descriptive levels of significance. Specifically, let $\hat{\underset{\sim}{\psi}}$ denote the least squares estimate of $\underset{\sim}{\psi}$ based on the full data set and $\hat{\underset{\sim}{\psi}}_{(-i)}$ the analogous estimate without the ith point. Then the generalized measure is defined by

$$D_i(\underset{\sim}{\psi}) = \frac{(\hat{\underset{\sim}{\psi}}_{(-i)} - \hat{\underset{\sim}{\psi}})' \underset{\sim}{B}^{-1} (\hat{\underset{\sim}{\psi}}_{(-i)} - \hat{\underset{\sim}{\psi}})}{qs^2} \tag{2}$$

where $s^2 = \underset{\sim}{R}'\underset{\sim}{R}/(n-p)$ and $\underset{\sim}{R} = (r_i)$ is the residual vector from the least squares analysis of equation (1). The magnitude of the distance between $\hat{\underset{\sim}{\psi}}$ and $\hat{\underset{\sim}{\psi}}_{(-i)}$ is assessed by comparing $D_i(\underset{\sim}{\psi})$ to the probability points of the central F-distribution with q and $n-p$ degrees of freedom.

Let $v_i = \underset{\sim i}{x}'(X'X)^{-1}\underset{\sim i}{x}$ where $\underset{\sim i}{x}'$ is the ith row of $\underset{\sim}{X}$ and $t_i = r_i s\sqrt{1-v_i}$ denotes the ith studentized residual. It can be shown that (see [2])

$$D_i(\underset{\sim}{\psi}) = \frac{t_i^2}{q} \frac{\underset{\sim i}{x}'(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{A}'\underset{\sim}{B}^{-1}\underset{\sim}{A}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim i}{x}}{1 - v_i} . \tag{3}$$

Further, letting $\underset{\sim p}{I}$ denote the $p \times p$ identity matrix, when $\underset{\sim}{A} = \underset{\sim p}{I}$ the generalized measure reduces to the basic measure,

$$D_i = D_i(\underset{\sim}{\beta}) = \frac{t_i^2 v_i}{p(1-v_i)} \tag{4}$$

$$= \frac{t_i^2 V(\hat{y}_i)}{pV(r_i)}$$

where $V(\hat{y}_i) = \sigma^2 v_i$ is the variance of the ith predicted value and $V(r_i) = \sigma^2(1-v_i)$ is the variance of the ith residual.

2. SUBSETS

We consider the important special case in which a subset, $\underset{\sim}{\beta}_2$, of q elements of $\underset{\sim}{\beta}$ is of interest. Without loss of generality we can take $\underset{\sim}{\beta}_2$ to be the last q elements of $\underset{\sim}{\beta} = (\underset{\sim}{\beta}_1, \underset{\sim}{\beta}_2)$ and, thus, $\underset{\sim}{A} = [\underset{\sim}{0}, \underset{\sim}{I}_q]$.

Let

$$(\underset{\sim}{X}'\underset{\sim}{X})^{-1} = \begin{pmatrix} \underset{\sim}{W}_{11} & \underset{\sim}{W}_{12} \\ \underset{\sim}{W}_{21} & \underset{\sim}{W}_{22} \end{pmatrix}$$

where $\underset{\sim}{W}_{22}$ is the $q \times q$ submatrix associated with $\underset{\sim}{\beta}_2$. Using this partition a little algebra will verify that

$$\underset{\sim}{B} = \underset{\sim}{W}_{22}$$

and

$$(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{A}'\underset{\sim}{B}^{-1}\underset{\sim}{A}(\underset{\sim}{X}'\underset{\sim}{X})^{-1} = (X'X)^{-1} - \begin{pmatrix} \underset{\sim}{K}^{-1} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{0} \end{pmatrix}$$

where $\underset{\sim}{K}^{-1} = \underset{\sim}{W}_{11} - \underset{\sim}{W}_{12}\underset{\sim}{W}_{22}^{-1}\underset{\sim}{W}_{21}$ is, apart from a proportionality constant, the covariance matrix of the least squares estimate of $\underset{\sim}{\beta}_1$ based only on the first $p-q$ independent variables. Substitution into equation (3) yields

$$D_i(\underset{\sim}{\beta}_2 | \underset{\sim}{\beta}_1) = \frac{t_i^2}{q} \frac{V(\hat{y}_i) - V_1(\hat{y}_i)}{V(r_i)}$$

$$= \frac{t_i^2}{q} \frac{v_i - v_{1,i}}{1 - v_i} \qquad (5)$$

where $V_1(\hat{y}_i) = \sigma^2 v_{1,i}$ denotes the variance of the ith predicted value from the regression on only the first $p-q$ variables. The notation "$D_i(\underset{\sim}{\beta}_2|\underset{\sim}{\beta}_1)$" is meant to remind us that we are considering distance based on a marginal confidence ellipsoid for $\hat{\underset{\sim}{\beta}}_2$, $\hat{\underset{\sim}{\beta}} = (\hat{\underset{\sim}{\beta}}_1, \hat{\underset{\sim}{\beta}}_2)$, from a fit to the full model. From equation (5) we see that the influence of an observation on a selected subset of $\hat{\underset{\sim}{\beta}}$ may be determined by combining the results of two separate regression.

Alternatively, equation (5) may be expressed as

$$D_i(\underset{\sim}{\beta}_2|\underset{\sim}{\beta}_1) = D_i(\underset{\sim}{\beta}) \cdot \left[1 - \frac{v_{1,i}}{v_i}\right] \frac{p}{q} .$$

Hence, $D_i(\underset{\sim}{\beta}_2|\underset{\sim}{\beta}_1) < D_i(\underset{\sim}{\beta})$ whenever $v_{1,i}/v_i < 1 - q/p$; also,

$$D_i(\underset{\sim}{\beta}_2|\underset{\sim}{\beta}_1) \leq D_i(\underset{\sim}{\beta}) \cdot \frac{p}{q} \qquad (6)$$

for all i. The latter inequality shows that it may not be necessary to use $D_i(\underset{\sim}{\beta}_2|\underset{\sim}{\beta}_1)$ directly when subsets of size q are of interest: If $D_i(\underset{\sim}{\beta})\frac{p}{q}$ is negligible then $D_i(\underset{\sim}{\beta}_2|\underset{\sim}{\beta}_1)$ must also be negligible over all subsets, $\underset{\sim}{\beta}_2$, of size q. This generalizes one of the results for the special case, q = 1, given in [2]. In short, if $pD_i(\underset{\sim}{\beta})$ is negligible for all i then no single observation has a serious influence on the least squares estimate of any subset of $\underset{\sim}{\beta}$. (Recall that the magnitude of $pD_i(\underset{\sim}{\beta})/q$ may be assessed by comparing it to the probability points of a central F-distribution with q and n-p degrees of freedom.)

It is also worth noting that if only the constant term is being ignored then $v_{1,i} = \frac{1}{n}$. Unless the constant term is of special interest it may be desirable to routinely use

$$D_i(\underset{\sim}{\beta}_2 \,|\, \text{constant term}) = \frac{t_i^2}{(p-1)} \; \frac{v_i - \frac{1}{n}}{(1 - v_i)} \qquad (7)$$

as an exploratory tool for isolating influential observations.

Once an influential observation has been detected using either equation (4) or (7) then the general form in equation (5) may be used to isolate its effects on the individual components of $\hat{\underset{\sim}{\beta}}$. It may be, for example, that the observation influences only one or two components of $\hat{\underset{\sim}{\beta}}$. The following example illustrates this.

3. EXAMPLE

Daniel and Wood [3] considered a set of data on the oxidation of ammonia to nitric acid. The original data set is from Brownlee [1] and consists of 21 observations with three possible explanatory variables. After a reasonably extensive analysis, Daniel and Wood decided that 4 observations (1, 3, 4 and 21) were outliers and that one of the explanatory variables was not needed. Their final model contained a linear and quadratic term for one explanatory variable, a linear term for the other and was based on 17 "valid" observations.

In this example we adopt the model

$$E(y_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 \qquad (8)$$

where y = stack loss

x_1 = air flow

and

x_2 = cooling water inlet temperature .

Also, only the observations that Daniel and Wood judged valid are considered. For ease of reference the data are given in Table 1. Notice that the original numbering of the observations has been maintained.

Table 2 gives the values of t_i, v_i, $D_i(\underset{\sim}{\beta})$ and $pD_i(\underset{\sim}{\beta})$ for the model in equation (8) and the observations listed in Table 1. Inspection of the first column in Table 2 shows that all observations appear to conform to the assumed model. However, inspection of the third column reveals that observation 2 is highly influential: Comparing $D_2(\underset{\sim}{\beta}) = 12.175$ to the probability points of the central F-distribution with 4 and 13 degrees of freedom shows that the removal of observation 2 would move the least squares estimate of $\underset{\sim}{\beta}$ to beyond the edge of the usual 99.95% elliptical confidence region for $\underset{\sim}{\beta}$ centered at $\hat{\underset{\sim}{\beta}}$. No other observation appears to exert much influence on $\hat{\underset{\sim}{\beta}}$.

Recall that the values of $pD_i(\underset{\sim}{\beta})$ listed in the fourth column of Table 2 are upper bounds for $D_i(\underset{\sim}{\beta}_2|\underset{\sim}{\beta}_1)$. Inspection of this column shows that observations 2 and 13 are the only ones that could have much of an influence on subsets of $\hat{\underset{\sim}{\beta}}$. In an effort to see how observation 2 influences subsets of $\hat{\underset{\sim}{\beta}}$ and if observation 13 influences any subset we consider next values of $D_2(\underset{\sim}{\beta}_2|\underset{\sim}{\beta}_1)$ and $D_{13}(\underset{\sim}{\beta}_2|\underset{\sim}{\beta}_1)$. Table 3 lists these values for selected subsets, $\hat{\underset{\sim}{\beta}}_2$. The results in Table 3 clearly indicate that observation 2 has a substantial influence on $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ but has little influence on $\hat{\beta}_3$. Also, observation 13 has little influence on the subsets investigated in Table 3.

The usual analysis seems to confirm these results: $\hat{\underset{\sim}{\beta}}_{(-2)} =$ (-56.79, 1.40, -.007, .601), $\hat{\underset{\sim}{\beta}}$ = (-15.41, -.069, .007, .528). Notice that the removal of observation 2 changes the sign of the estimate of β_3.

Also, in the complete data set the partial F-statistics for $\hat{\beta}_2$ and $\hat{\beta}_3$ are significant while those for $\hat{\beta}_0$ and $\hat{\beta}_1$ are not significant at the usual levels. With observation 2 deleted the partial F-statistics for $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ are all less than one while the statistic for $\hat{\beta}_3$ remains large. It appears that for the final data set of Daniel and Wood the quadratic term is needed to model a single observation. The use of such an influential observation without an independent verification of its authenticity or a well-grounded firm belief in the model does not seem to be sound practice.

The reason for the importance of observation 2 can be obtained from the second column in Table 2, $v_2 = 0.993$. This large value for v_2 suggests that observation 2 corresponds to outlying values in the independent variables. Inspection of the data in Table 1 show this to be the case. The largest values of x_1 and x_2 both occur at observation 2. Generally, it can be shown that the observation corresponding to $\max v_i$ must lie on the boundary of the convex hull of the design points (i.e. the rows of $\underset{\sim}{X}$). In this example, observation 2 lies on the boundary of the convex hull of the design points and is considerably removed from the bulk of the design points.

4. COMMENT

The inequality in (6) which was stated to hold over all subsets of q elements of $\hat{\underset{\sim}{\beta}}$ actually remains true over all possible sets of q linearly independent combination of the elements of $\hat{\underset{\sim}{\beta}}$. This is easily shown as follows:

Assume that interest is in the q linearly independent combinations, $\underset{\sim}{\psi}_2$, specified by $\underset{\sim}{\psi}_2 = \underset{\sim}{A}_2\underset{\sim}{\beta}$ where $\underset{\sim}{A}_2$ is a $q \times p$ rank q matrix. Let $\underset{\sim}{A}_1$ denote a $p-q \times p$ matrix such that $\underset{\sim}{A}' = [\underset{\sim}{A}_1', \underset{\sim}{A}_2']$ is of full rank, p, and define $\underset{\sim}{\psi}_1 = \underset{\sim}{A}_1\underset{\sim}{\beta}$ and $\underset{\sim}{\psi}' = (\underset{\sim}{\psi}_1', \underset{\sim}{\psi}_2')$. Consider the transformed model,

$$\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{A}^{-1}\underset{\sim}{\psi} + \varepsilon.$$

Since $D_i(\underset{\sim}{\beta})$ is invariant under nonsingular linear transformations of the columns of $\underset{\sim}{X}$, we have

$$D_i(\underset{\sim}{\beta}) = D_i(\underset{\sim}{\psi})$$

for all $p \times p$ nonsingular matrices, $\underset{\sim}{A}$. We now apply the discussion of Section 2 to the subset $\underset{\sim}{\psi}_2$ of the transformed model and obtain

$$D_i(\underset{\sim}{\psi}_2 | \underset{\sim}{\psi}_1) \leq D_i(\underset{\sim}{\psi}) \frac{p}{q} = D_i(\underset{\sim}{\beta}) \frac{p}{q}.$$

Thus, if $D_i(\underset{\sim}{\beta}) \frac{p}{q}$ is small the ith observation has a negligible influence on all possible sets of q linearly independent combinations of the elements of $\hat{\underset{\sim}{\beta}}$.

REFERENCES

[1] Brownlee, K.A., (1965). Statistical Theory and Methodology in Science and Engineering. Wiley, New York.

[2] Cook, R.D., (1977). The detection of influential observations in linear regression. Technometrics, 19, 15-18.

[3] Daniel, D. and Wood, F.S., (1971). Fitting Equations to Data. Wiley-Interscience, New York.

TABLE 1

Data on the Oxidation of
Ammonia to Nitric Acid

Observation Number	Air Flow x_1	(Air Flow)^2 x_1^2	Cooling Water Inlet Temperature x_2	Stack Loss y
2	80	6400	27	37
5	62	3844	22	18
6	62	3844	23	18
7	62	3844	24	19
8	62	3844	24	20
9	58	3364	23	15
10	58	3364	18	14
11	58	3364	18	14
12	58	3364	17	13
13	58	3364	18	11
14	58	3364	19	12
15	50	2500	18	8
16	50	2500	18	7
17	50	2500	19	8
18	50	2500	19	8
19	50	2500	20	9
20	56	3136	20	15

TABLE 2

Values of t_i, v_i, $D_i(\underset{\sim}{\beta})$ and $pD_i(\underset{\sim}{\beta})$ for the data of Table 1.

Observation	t_i	v_i	$D_i(\underset{\sim}{\beta})$	$pD_i(\underset{\sim}{\beta})$
2	0.57	0.993	12.175	48.698
5	-0.12	0.131	0.001	0.002
6	-0.64	0.164	0.021	0.082
7	-0.18	0.242	0.003	0.011
8	0.84	0.242	0.056	0.223
9	-0.66	0.208	0.028	0.113
10	0.96	0.179	0.051	0.202
11	0.96	0.179	0.051	0.202
12	0.53	0.280	0.028	0.111
13	-1.98	0.179	0.213	0.852
14	-1.41	0.113	0.068	0.272
15	0.32	0.193	0.006	0.024
16	-0.67	0.193	0.027	0.108
17	-0.21	0.197	0.003	0.010
18	-0.21	0.197	0.003	0.010
19	0.27	0.237	0.006	0.022
20	2.16	0.070	0.088	0.352

TABLE 3

Values of $D_2(\underset{\sim}{\beta}_2|\underset{\sim}{\beta}_1)$ and $D_{13}(\underset{\sim}{\beta}_2|\underset{\sim}{\beta}_1)$ for Selected Subsets, $\underset{\sim}{\beta}_2$.

$\underset{\sim}{\beta}_2$	$D_2(\underset{\sim}{\beta}_2\|\underset{\sim}{\beta}_1)$	$D_{13}(\underset{\sim}{\beta}_2\|\underset{\sim}{\beta}_1)$
$(\beta_0, \beta_1, \beta_2, \beta_3)$	12.175	0.213
(β_0)	13.428	0.324
(β_1)	13.670	0.100
(β_2)	17.983	0.057
(β_3)	0.235	0.463
(β_0, β_3)	6.717	0.425
(β_1, β_3)	6.922	0.286
(β_2, β_3)	8.994	0.286
$(\beta_0, \beta_1, \beta_2)$	16.134	0.283